



ALEXA-Seq
(www.AlexaPlatform.org)

User Manual (v.1.5)

23 April 2010



Table of Contents

Table of Contents	2
Authors	3
Citation	3
License	3
Acknowledgements	3
Affiliations	3
Introduction	3
Before Starting	4
System requirements	4
Warnings	4
Example data	5
Dependencies	5
Obtaining ALEXA-Seq code	7
Installation of ALEXA-Seq code	7
Script and Module Locations	8
Configuration	8
1. Analysis configuration	8
2. Project configuration	9
3. Installation of ALEXA-Seq annotation database	9
A. Download and install an existing database (recommended)	9
B. Request the database	10
C. Create new ALEXA-Seq database	10
Analysis	11
Command creation	11
0.) Install the Ensembl API	11
1.) Create target directories	11
2.) Import raw read data	11
3.) Basic data statistics	12
4.) Generate read fasta files	12
5.) Mapping	12
6.) Read assignment (aka 'parsing')	12
7.) Read assignment summary	13
8.) Expression calculations	13
9.) Summarize expression statistics	13
10.) Regenerate expression values	13
11.) Create custom UCSC track files to visualize data	13
12.) Calculate Differential Expression (DE) of all features	14
13.) Calculate Alternative Expression (AE) of all features	14
14.) Populate ALEXA-Seq data viewer	14
15.) Index all gene records using the Xapain-Omega utility	14
Result file locations	15

Authors

ALEXA-Seq is the work of Malachi Griffith and Marco A. Marra.

Citation

Malachi Griffith, Obi L. Griffith, Ryan D. Morin, Michelle J. Tang, Ying-Chen Hou, Trevor J. Pugh, Rodrigo Goya, Jill Mwenifumbo, Suganthi Chittaranjan, Adrian Ally, Jennifer K. Asano, Susanna Y. Chan, Haiyan I. Li, Helen McDonald, Kevin Teague, Yongjun Zhao, Thomas Zeng, Allen Delaney, Martin Hirst, Gregg B. Morin, Steven J. M. Jones, Isabella T. Tai, Marco A. Marra*. *Alternative expression analysis by RNA sequencing*. Pending publication.

License

ALEXA-Seq is open source and available for public use under the terms of the GNU General Public License, version 3. For details refer to:

<http://www.gnu.org/licenses/gpl.txt>

Acknowledgements

We are grateful for funding provided by the following organizations: the University of British Columbia, Faculty of Graduate Studies and Faculty of Medicine, the Michael Smith Foundation for Health Research, the Natural Sciences and Engineering Research Council, Genome British Columbia, the National Cancer Institute of Canada and the Terry Fox Foundation.

Affiliations

British Columbia Cancer Agency - Genome Sciences Centre

University of British Columbia - Faculty of Medicine – Department of Medical Genetics

Introduction

This manual provides instructions to assist in the completion of an ALEXA-Seq analysis (Alternative Expression Analysis by massively parallel RNA sequencing). The purpose of the ALEXA-Seq method is to comprehensively profile the expression, differential expression and alternative expression of a transcriptome and compare transcript specific expression events between two or more conditions of interest. Novel transcript discovery is also facilitated. These analyses require as input, paired-end RNA-Seq (aka WTSS) data from a high throughput sequencing device (such as an Illumina GAI). For examples of the output of the analysis please refer to our website. Briefly, the output consists of expression, differential expression and alternative expression values for transcripts as well as their component exons, junctions, boundaries, introns, etc.

In order to complete an ALEXA-Seq analysis you will need to download the source code and other resources from our website. The website also contains example data, and the results of many ALEXA-Seq analyses performed by us as part of various collaborations. As a department of the BC Cancer Agency, our focus in these collaborations has primarily been the analysis of cancer samples. However, the method described below was designed to work for any comparison between paired or grouped conditions. Although we have primarily analyzed human and mouse data, the method should also work for any species currently annotated by Ensembl. For further details please visit our website: www.AlexaPlatform.org

Before Starting

Before you can run the ALEXA-Seq analysis, you will need to ensure that your system meets certain requirements. We make use of numerous existing bioinformatics tools. Many of these are commonly used, but you should check the list of dependencies below to ensure that all necessary tools are installed. If you have problems that are not covered by this manual, please contact us. Contact information is available on our website: www.AlexaPlatform.org

System requirements

The ALEXA-Seq pipeline is designed to run in a 32- or 64-bit Linux system. We use RHEL 4 and CentOS 5, but any distribution should work fine. Due to the extremely large datasets generated by next-generation sequencing devices you will require considerable CPU and storage resources to perform an ALEXA-Seq analysis. We used about ~20-100 Gb of disk space to process each of the datasets listed in the results section of our website (www.AlexaPlatform.org). Most of this is needed only temporarily and the final result should take 5-10 Gb of storage space.

We also had access to a Beowulf style CPU cluster of ~1000 nodes. Although, not technically needed to perform ALEXA-Seq analyses, it is highly recommended to increase the rate of data processing. Detailed instructions will be provided for submission of jobs to a CPU cluster. However, since configuration of the CPU cluster will likely differ at your centre, these steps may need to be modified slightly. The jobs themselves should not change but the steps involved in submitting those jobs to your cluster may vary from ours.

Finally, in order to create your own ALEXA-Seq data viewer to help visualize and distribute data, you will need access to a web server (we use APACHE2).

Warnings

You will require at least a basic familiarity with Linux and Perl to perform the ALEXA-Seq analysis. Furthermore, certain aspects of this pipeline are extremely computationally intensive. I have included scripts to assist in the creation of parallel jobs to be run on a cluster of computers. If you do not have access to such computer resources, this analysis may not be practical for large datasets.

Example data

Example datasets can be downloaded from our website as they become available.

http://www.alexaplatform.org/alex_seq/results.htm

Dependencies

You will need the following components to be installed. Some of these may be included with your Linux distribution by default. Others will need to be installed. For a detailed walkthrough of the installation procedure, including all dependencies refer to the separate ALEXA-Seq Linux installation manual included with the source code.

Cairo

<http://cairographics.org/>

In order to dynamically produce graphics in R you will need to install both the 'cairo' and 'pixmap' packages for Linux.

R – The R Project for Statistical Computing

<http://www.r-project.org/>

Once R is installed, begin a session and install the following additional packages:

'RColorBrewer', 'Cairo'

Bioconductor (biocLite)

<http://www.bioconductor.org/>

Bioconductor is used for certain data processing and graphing functions. The biocLite version contains all necessary packages.

Perl

<http://www.perl.org/>

Perl should already be installed with Linux. Both Perl 5.6.x and 5.8.x should work.

BioPerl

<http://www.bioperl.org/Core/Latest/index.shtml>

BioPerl is used by the Ensembl API. BioPerl should be installed into the 'ensembl_api' sub-directory of alexa_seq. This installation will be handled automatically by the pipeline.

Berkeley DB Perl Module

<http://search.cpan.org/dist/BerkeleyDB/BerkeleyDB.pod.P>

Needs to be installed for some components of ALEXA-Seq to work. Ideally it would be installed globally in your system so that Perl knows where to find it. If this is not possible, instructions are provided in INSTALL.txt with the source code for setting the appropriate environment variables.

Ensembl API

<http://ensembl.org/info/data/api.html>

Updates should be installed into the 'ensembl_api' sub-directory of alexa_seq. This installation will be handled automatically by the pipeline.

MySQL – Community Server

<http://dev.mysql.com/downloads/>

MySQL may be included with your Linux distribution. ALEXA-Seq pipeline uses mysql databases to access Ensembl data via the Ensembl API as well to store custom annotation data. We have been using MySQL version 4.1.

MySQL DBI/DBD

<http://search.cpan.org/dist/DBI/>

<http://search.cpan.org/dist/DBD-mysql/>

DBI/DBD may be included with your Linux distribution by default. These are two Perl modules that allow Perl to interact directly with MySQL database.

BLAST

<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/>

Download the 32- or 64-bit version as needed.

To ensure compatibility of sequence databases, use the same version as us: 2.2.18. Otherwise you will have to update the databases (blast provides a utility for this) or create new databases.

mdust

<http://compbio.dfci.harvard.edu/tqi/software/>

The version we use in the pipeline is included with the ALEXA-Seq source code package in the 'external_tools' sub-directory. Unpack and compile this tool. Note the location of the directory containing the 'mdust' binary. This information will be needed when setting up your ALEXA-Seq configuration file.

Xapian-Omega

<http://xapian.org/>

Used to index results created by the ALEXA-Seq pipeline. This tool will need to be installed somewhere that is accessible by your web server. It will also need permission to run as an executable process on the web server.

Obtaining ALEXA-Seq code

There are several options for getting the ALEXA-Seq code.

A.) From our website

The source code package as well as ALEXA-Seq annotation databases can be downloaded from here. The code is available

<http://www.alexaplatform.org/alexaseq/downloads.htm>

B.) From SourceForge.net

<https://sourceforge.net/projects/alexaseq/>

C.) From our public subversion repository

If you have subversion ('svn') installed, you can check out the code directly from our subversion repository. For example, execute the following from a command prompt

```
svn co https://svn.bcgsc.ca/public/ALEXA_Seq/tags/ALEXA_Seq_v1.5 ALEXA_Seq_v1.5
```

Installation of ALEXA-Seq code

The following is a general overview of the installation procedure, and assumes all dependencies (p 5-6) are already installed. For a detailed walkthrough of the installation procedure, including all dependencies refer to the ALEXA-Seq Linux installation manual.

Before proceeding, download and unpack the ALEXA code base. For example:

```
mkdir /home/user/ALEXA/  
cp ALEXA_Seq_v.1.5.tar.gz /home/user/ALEXA/  
cd /home/user/ALEXA/  
gunzip ALEXA_Seq_v.1.5.tar.gz  
tar -xvf ALEXA_Seq_v.1.5.tar
```

Unpacking the code is not necessary if you obtained it from our svn repository.

Once the code is installed. Open the INSTALL.txt file and copy the indicated section of environment variables to your .bashrc file (or an equivalent shell parameters file that is automatically loaded on log in). These variables should be edited to reflect your own specific file paths and install directories. Finally, once this is done, log out of your session and log in again to make sure the environment variables take effect.

Script and Module Locations

The root/reference directory for all scripts is: '~/.alexa_seq' (where ~ is wherever the code was unpacked). Many of the scripts described below make use of methods that we have written and stored in utility modules. The details of these functions are beyond the scope of this document. These Perl modules follow POD (plain old documentation) format and are stored in: '~/.alexa_seq/utilities'. Each script can be run without arguments to display a list of instructions.

Configuration

Before starting your analysis you will need to track down basic information about your computer system as well as specific info relating to the project. A 'project' relates to a biological question. Many examples of such projects that have already been analyzed by ALEXA-Seq are provided at our website:

http://www.alexaplatform.org/alexa_seq/results.htm

For example, a project could consist of a tumour versus normal comparison, or treated versus untreated cell lines, etc. Each project generally has two or more sequence 'libraries' corresponding to these conditions. Each sequence library may itself consist of multiple 'lanes' of paired-end sequence data that were generated as one or more runs of a high-throughput sequencing device. Each project generally also has at least one 'comparison' defined (e.g. tumour vs. normal).

1. Analysis configuration

For tidiness, you should have three main directories for ALEXA-Seq processing. Briefly, these contain the code itself, the sequence databases, and the analysis files. For example these directories could be:

```
/home/user/ALEXA/alexa_seq/  
/home/user/ALEXA/sequence_databases/  
/home/user/ALEXA/alexa_seq_analysis/
```

Other folders you may wish to create:

```
/home/user/ALEXA/perl_storables/ (for BerkeleyDB perl storables?)  
/home/user/ALEXA/www/ (for web files in case you can't write directly to your web server)  
/home/user/ALEXA/commands/ (for .commands files for each project)  
/home/user/ALEXA/config_files/ (for config_files for each project)
```

Before performing your first analysis, you will need to create an updated analysis configuration file. This file stores basic system specific parameters, mostly paths to directories or binaries.

Create this file by starting with the example file provided:

```
cp ~/.alexa_seq/config_files/examples/ALEXA_Seq_PIPELINE.conf
```


~/config_files/ALEXA_Seq_PIPELINE.conf

Edit all values in the config files to reflect your own file locations.

The validity of values entered in the configuration files will be tested in the first step of the analysis (by createAnalysisCommands.pl - see below).

2. Project configuration

Before starting analysis for a new project (and for the first analysis you run), you will also need to create a project configuration file. This file stores information about the data to be analyzed. Specifically, you must define each 'LANE', 'LIBRARY', and 'COMPARISON'. For example, you will need to determine the paths to your input sequence data. Also included in this file are configuration values that might be project specific (e.g. the species being analyzed, the genome version to use, etc.).

Create this file by starting with the example file provided here:

~/alexa_seq/config_files/examples/ALEXA_Seq_<ExampleProject>.conf

Several example project configurations are provided. The validity of values entered in this configuration file will be tested in the first step of the analysis.

3. Installation of ALEXA-Seq annotation database

You will need an ALEXA-Seq annotation database to perform the analysis. This database defines all the canonical, alternative and hypothetical sequence features that will be considered in the analysis. The database is specific to a particular species and genome build of that species. You have three options for obtaining these databases. Instructions for each of these options are as follows:

A. Download and install an existing database (recommended)

Pre-computed databases are provided here:

ftp://ftp03.bcgsc.ca/public/ALEXA/alexa_seq/

Each of these is described here:

http://www.alexaplatform.org/alexa_seq/downloads.htm

Most of the analyses described on our website used the Human build 'hs_53_36o' (NCBI Build 36/UCSS hg18) but databases for several other species and builds are available.

Install as follows using 'hs_53_36o' as an example:

First log into your mysql server and create the target database

```
mysql
CREATE DATABASE ALEXA_hs_53_36o
exit
```

Then download and install the database:

```
cd /home/user/sequence_databases/
wget ftp://ftp03.bcgsc.ca/public/ALEXA/alexa_seq/hs_53_36o.tar.gz
gunzip hs_53_36o.tar.gz
tar -xvf hs_53_36o.tar
cd hs_53_36o/alexa_db/
gunzip ALEXA_hs_53_36o.tables.tar.gz
tar -xvf ALEXA_hs_53_36o.tables.tar
~/alexa_seq/sql/restoreAlexaDb.pl
```

B. Request the database

If your analysis involves a species we have not included as a download, you can request that we add it by contacting us through our website (www.AlexaPlatform.org). We currently have ALEXA-Seq annotation databases for: Chicken, Chimp, Fly, Human, Mouse, Rat, Yeast, and Zebrafish.

C. Create new ALEXA-Seq database

If you wish to create a complete new annotation database you can use the following tool to create the necessary instructions:

```
~/alexa_seq/createAnnotationCommands.pl
```

You will be asked to supply two configuration files. One was created above. For the other, use the following file as an example:

```
~/alexa_seq/config_files/examples/ALEXA_Seq_dr_57_8c.conf
```

This script will create a new '.commands' file containing detailed instructions for building your own ALEXA-Seq annotation database.

A detailed description of the database schema is provided here:

http://www.alexaplatform.org/alexa_seq/data/ALEXA_Seq_Schema_Description.htm

Analysis

Command creation

Once you have created or updated your system and project configuration files as described above, execute the following to create the analysis commands:

```
cd ~/alexa_seq/  
./createAnalysisCommands.pl  
--alexa_seq_config_file=config_files/examples/ALEXA_Seq_PIPELINE.conf  
--project_config_file=config_files/examples/ALEXA_Seq_<ExampleProject>.conf  
--commands_file=ALEXA_Seq_ExampleProject.commands
```

Open the '.commands' file created by this script and follow the instructions. The tasks that will be performed are briefly described below. *Each step described below corresponds to a step in the .commands file with the same number.*

0.) Install the EnsEMBL API and BioPerl

The version to use is specified in the configuration file. It will only be installed if not already present.

1.) Create target directories

All directories need for the analysis and creation of results files, statistics, figures etc. will be automatically created.

2.) Import raw read data

This step will concatenate (if necessary) the *_seq.txt or *_qseq.txt files from a source directory to create a combined raw seq data file. There may be many tile files corresponding to a single lane, or they may already be joined together. These files are expected to be named as follows:

```
s_1_1_0001_qseq.txt.bz2  
i.e. s_<Lane>_<Read-1-or-2>_Tile_<seq-or-qseq>.txt.<gz-or-bz2>)
```

Note: compressing seq or qseq files is not required but recommended to save storage space.

The format of sequence data in the raw .seq or .qseq files is as follows:

Lane	Tile	X-coord	Y-Coord	Sequence
------	------	---------	---------	----------

In this format, the sequence for read1 and read2 is pasted together for paired reads. Bases which could not be resolved are represented by a '.' and will be converted to N's. During this step, the complexity of each read will be determined by 'mdust' and various statistics pertaining to each lane will be summarized (total read counts, low complexity reads, poor quality reads, etc.). During this step, it is also possible to trim reads if there was a problem with the quality of read ends.

3.) Basic data statistics

Gather basic info about each lane, library and comparison defined in the Project configuration file. Also get number of quality reads, the average read length, and the overall tag redundancy of each library.

3-A.) Gather info about each lane of data

Imported from the project configuration file.

3-B.) Gather info about each library

Imported from the project configuration file.

3-C.) Gather info about each comparison

Imported from the project configuration file.

3-D.) Generate statistics for each library

This step will determine: the number of quality read counts in each lane, the average read length of the library, and the overall tag redundancy of each library. If the library consists of a mixture of read lengths (not advisable) it may be useful to know the average read length.

4.) Generate read fasta files

One fasta file will be created for each lane of data after filtering out reads that are poor quality (too many ambiguous bases, i.e. N's), low complexity (e.g. polyA reads), or where both reads of a pair are identical (library artifacts).

5.) Mapping

Reads will be mapped to a database of repeat elements, transcripts, known exon junctions and boundaries, hypothetical exons junctions and boundaries, and intronic and intergenic regions of the genome. Where possible, read-pairing information is used to resolve ambiguously mapped reads but if only one read of a pair can be mapped this read will still be retained (i.e. paired mapping is not required). With default parameters reads are mapped with up to 3 mismatches and 1 gap. Sub-string alignments are also allowed. A perfect alignment of 60% of the read length will be allowed (again assuming default parameters).

6.) Read assignment (aka 'parsing')

Based on the alignments, reads will now be assigned to their most probable source. If possible each read is unambiguously assigned to a repeat element, known transcript, novel exon junction, novel exon boundary, intron, or intergenic region. At this stage, the apparent distance between reads of a pair is determined as well as other statistics pertaining to mapping efficiency. Reads that can not be assigned to a repeat or human genome or transcriptome sequence are retained but are marked as 'Unassigned' and excluded from downstream analysis.

7.) Read assignment summary

The number of reads assigned to each category/read class for each library will be determined.

8.) Expression calculations

Using the coordinate information in the ALEXA-Seq annotation database in combination with the mapping results generated above, the expression of 13 types of sequence features will be determined. These features consist of: Genes, transcripts, exon regions, exon junctions (known and novel), exon boundaries (known and novel), introns, active intronic regions, silent intronic regions, intergenic regions, active intergenic regions and silent intergenic regions. Refer to the manuscript methods for further details on the annotation of these features. The result of this analysis for human is an expression value for ~4 million sequence features.

Transcript specific expression for all known Ensembl transcripts is calculated by using only those exon regions and exon junctions that are unique to each transcript.

For each feature, several expression metrics are calculated. These include cumulative coverage, average base coverage (cumulative coverage divided by the length of the sequence feature), the percentage of bases of a feature covered at 1x or greater, etc. See the schema description for a detailed explanation of all expression measures.

http://www.alexaplatform.org/alexaplatform/seq/data/ALEXA_Seq_Schema_Description.htm

The expression of exon junction and boundary features are systematically lower than those for exon regions. This is due to an inherent mapping disadvantage for these sequences (see manuscript for discussion). For this reason, these values are empirically adjusted to compensate for the disadvantage.

9.) Summarize expression statistics

In this step, various expression statistics and graphs are generated. For each library, gene specific expression cutoff values are determined by examination of intronic and intergenic noise levels (see manuscript for details). The number of features of each type that are expressed above background is then determined. The overall sequence coverage of genes is also determined in this step.

10.) Regenerate expression values

Using the expression values determined above, the junction/boundary correction factors and gene-by-gene cutoffs values are now used to re-evaluate the expression of all features. To accomplish this, steps 8-9 are repeated.

11.) Create custom UCSC track files to visualize data

Expression values will be used to create custom UCSC track files (a mixture of GFF and wig tracks) to display all features expressed above background and the base level sequence coverage of the genome. Similarly a wig track displaying base level differential expression will also be created in this step for each comparison you define.

12.) Calculate Differential Expression (DE) of all features

Differential expression of all features (Genes, Exons, Junctions, Boundaries, etc.) will be assessed as the log2 difference in expression level for each comparison you define. Each differential expression value will also be associated with a p-value.

13.) Calculate Alternative Expression (AE) of all features

Alternative expression of all features (Genes, Exons, Junctions, Boundaries, etc.) will be assessed by calculating splicing index (SI) values, reciprocity index (RI), and percent feature contribution (PFC) values for each comparison you define (see manuscript for details).

14.) Populate ALEXA-Seq data viewer

To help summarize and visualize the expression of known and novel isoforms as well as differential and alternative gene expression between conditions of interest, all expression, differential expression and alternative expression results will be imported into the ALEXA-Seq data viewer (essentially a dynamically generate web interface). This viewer also facilitates data sharing with collaborators, design of validation experiments, and interpretation of the results.

In addition to these results, detailed summaries of the characteristics of each library are also produced. For example:

<http://www.alexaplatform.org/alexaseq/Morgen/MM0472.htm>

Several examples of data sets processed and displayed in the ALEXA-Seq data viewer are available at our website:

<http://www.alexaplatform.org/alexaseq/results.htm>

15.) Index all gene records using the Xapian-Omega utility

For every project, a summary is available for every gene. To allow searching for arbitrary genes, the results will be also indexed with the tool 'Xapian-Omega' in this step.

An example of this search functionality can be found here:

<http://www.bcgsc.ca/xapian-search/omega>

Result file locations

The following is a brief description of files created during the ALEXA-Seq analysis and their relative storage location (all files are created automatically during the analysis).

In the following descriptions ‘\$analysis_dir’ refers to the base path where all the results are stored (e.g. /home/user/alexa_seq_analysis/). This parameter is set at the beginning of the analysis in the ALEXA-Seq configuration file. Similarly, ‘\$project_name’ is the name of the project defined in the configuration file and ‘\$library_id’ is a variable used to denote each of the libraries being processed for that project.

Batch files

All command files to be executed for a particular project are stored as follows:

[\\$analysis_dir/batch_jobs/\\$project_name/](#)

Raw sequence data

Raw sequence files imported from the high-throughput sequencing device are stored as follows:

[\\$analysis_dir/raw_seq_data/\\$library_id/](#)

Fasta sequence files

Fasta files containing all reads passing basic quality filters are stored for each lane of each library here:

[\\$analysis_dir/fastq_seq_data/\\$library_id/](#)

Read record files

These files store the read sequences of each pair on a single line along with basic quality metrics for the read and its current assignment status. All reads start as ‘Unassigned’ and if possible are assigned to transcripts, junctions, introns, etc.

[\\$analysis_dir/read_records/\\$library_id/](#)

Mapping results files

Within the ‘read_records’ directory, mapping results files are stored in sub-directories. For convenience, of downstream analysis, individual mapping results are provided for each target sequence type (repeats, transcripts, junctions, boundaries, introns, and intergenic regions).

[\\$analysis_dir/read_records/\\$library_id/\\$sequence_type/](#)

Figures and statistics

A large number of statistics, summary files and figures are generated for each lane of data, library and comparison.

Summary of annotations used for the analysis:

[\\$analysis_dir/figures_and_stats/Generic/](#)

Basic lane-by-lane library quality statistics:

[\\$analysis_dir/read_records/\\$library_id/Summary/](#)

Feature expression, expression correlations, read assignments, average coverage values, library statistics, etc.:

[\\$analysis_dir/figures_and_stats/\\$library_id/Expression_v*/](#)

Differential expression results (by sequence feature type):

[\\$analysis_dir/figures_and_stats/DE/\\$project_name/\\$feature_type/](#)

Alternative expression results (by sequence feature type):

[\\$analysis_dir/figures_and_stats/SE/\\$project_name/\\$feature_type/](#)

Log files

Log files from the ALEXA-Seq analysis:

[\\$analysis_dir/logs/\\$library_id/](#)

Temporary files

Working directory for creation of temporary files during data processing:

[\\$analysis_dir/temp/](#)

Temporary storage for ALEXA-Seq data viewer files

[\\$analysis_dir/temp/\\$project_name/](#)